LECTURE 12 | المحاضرة 12

Reinforcement Learning in Energy Management

التعلم المعزز في إدارة الطاقة

Sequential decision-making for batteries, microgrids, demand response, and grid control

اتخاذ القرار المتتابع للبطاريات والشبكات المصغرة والاستجابة للطلب والتحكم بالشبكة

Prepared by | إعداد: Dr.-Ing. Aouss Gabash

Learning Outcomes

مخرجات التعلم

1. Design and evaluate an AI load-forecasting workflow.
   تصميم وتقييم سير عمل للتنبؤ الذكي بالأحمال.
2. Interpret the governing mathematical or algorithmic principles.
   تفسير المبادئ الرياضية أو الخوارزمية الحاكمة.
3. Connect the topic to practical electrical-power-engineering applications.
   ربط الموضوع بتطبيقات عملية في هندسة الطاقة الكهربائية.

Prerequisites

المتطلبات السابقة

Time-series analysis, regression, and load characteristics.

تحليل السلاسل الزمنية والانحدار وخصائص الأحمال.

AI-Based Load Forecasting | التنبؤ بالأحمال القائم على الذكاء الاصطناعي

🎯 Learning Objectives

- Formulate energy management as a Markov decision process.
- Define state, action, reward, policy, and return.
- Derive Bellman equations and Q-learning updates.
- Explain DQN, actor-critic, PPO, DDPG, and SAC.
- Design a reward for battery scheduling.
- Recognize safety and generalization challenges.

🎯 أهداف التعلم

- صياغة إدارة الطاقة كعملية قرار ماركوفية.
- تعريف الحالة والفعل والمكافأة والسياسة والعائد.
- اشتقاق معادلات بيلمان وQ-learning.
- فهم DQN وActor-Critic وPPO وDDPG وSAC.
- تصميم مكافأة لجدولة البطارية.
- فهم الأمان والتعميم.

1. Reinforcement Learning Loop

State s_t → Agent selects a_t → Environment →

Reward r_t and next state

> The agent is not given the correct action for every state. It discovers a policy through interaction.

1. حلقة التعلم المعزز

يراقب الوكيل الحالة ويختار فعلًا، ثم تنتقل البيئة إلى حالة جديدة وتعيد مكافأة. ومن تكرار هذه العملية يتعلم الوكيل سياسة القرار.

2. Markov Decision Process

$$\mathrm{MDP} = (\mathcal{S}, \mathcal{A}, P, R, \gamma)$$

State
Information for decision-making.

Action
Available control decision.

Transition
Evolution after the action.

Reward
Immediate decision quality.

Discount
Importance of future reward.

Policy
State-to-action mapping.

2. عملية قرار ماركوفية

تتكون من الحالات والأفعال وديناميكا الانتقال والمكافأة ومعامل الخصم والسياسة.

3. Markov Property

$$P(s_{t+1} \mid s_0, a_0, \dots, s_t, a_t) = P(s_{t+1} \mid s_t, a_t)$$

For battery control, SOC alone may be insufficient; time, load, price, renewable output, and forecasts may also be needed.

3. خاصية ماركوف

يجب أن تحتوي الحالة على المعلومات المهمة للمستقبل، ولذلك نضيف إلى SOC الزمن والسعر والحمل والتوقعات.

4. Return and Discount

$$G_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots$$

$$G_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}$$

A small γ emphasizes immediate savings; γ near one emphasizes long-term operation.

4. العائد والخصم

العائد هو مجموع المكافآت الحالية والمستقبلية بعد خصمها، ويحدد γ مقدار اهتمام الوكيل بالمستقبل.

5. Value Functions

$$V^{\pi}(s) = \mathbb{E}_{\pi}[G_t \mid s_t = s]$$

$$Q^{\pi}(s,a) = \mathbb{E}_{\pi}[G_t \mid s_t = s,\ a_t = a]$$

V evaluates a state; Q evaluates a particular action in that state.

5. دوال القيمة

تقيس V جودة الحالة، بينما تقيس Q جودة اختيار فعل معين ضمن تلك الحالة.

6. Bellman Optimality

$$Q^*(s,a) = \mathbb{E}\left[r + \gamma \max_{a'} Q^*(s',a')\right]$$

> A long-horizon problem becomes an immediate reward plus the best value of the next state.

6. أمثلية بيلمان

تربط قيمة القرار الحالي بالمكافأة الفورية وأفضل قيمة مستقبلية، وهي الأساس الرياضي لخوارزميات القيمة.

7. Q-Learning

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha\Big[r_{t+1} + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t)\Big]$$

$$\delta_t = r_{t+1} + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t)$$

7. خوارزمية Q-Learning

تُحدّث قيمة الفعل باستخدام خطأ الفرق الزمني بين القيمة الحالية والهدف الجديد.

8. Exploration and Exploitation

$$\begin{cases} \text{choose a random action}, & \text{with probability } \varepsilon, \\ \arg\max_a Q(s, a), & \text{otherwise.} \end{cases}$$

Too little exploration can trap the agent in a poor policy. Excessive exploration wastes experience and may violate limits.

8. الاستكشاف والاستغلال

توازن سياسة ε-greedy بين تجربة أفعال جديدة واستخدام أفضل فعل معروف.

9. Deep Q-Network

DQN approximates Q(s,a) using a neural network and is suited to discrete actions.

$$L(\theta) = \mathbb{E}\left[\left(r + \gamma \max_{a'} Q(s', a'; \theta^-) - Q(s, a; \theta)\right)^2\right]$$

Experience replay
Random mini-batches reduce correlation.

Target network
Stabilizes the learning target.

9. شبكة DQN

تستخدم DQN شبكة عصبية لتقريب Q، وتناسب الأفعال المنفصلة مثل الشحن أو التوقف أو التفريغ.

10. Actor-Critic Methods

$$\nabla_\theta J(\theta) = \mathbb{E}[\nabla_\theta \log \pi_\theta(a \mid s) \cdot A(s, a)]$$

Actor
Produces an action or probability distribution.

Critic
Estimates value and guides the actor.

10. أساليب Actor-Critic

يتعلم Actor السياسة مباشرة، بينما يقيم Critic جودة الأفعال ويوجه تحديث السياسة.

11. Algorithm Selection

Algorithm	Action space	Typical energy use
DQN	Discrete	Charge / idle / discharge
DDPG / TD3	Continuous	Continuous storage or inverter power
PPO	Discrete or continuous	General robust control policy
SAC	Continuous	Exploratory storage and microgrid control

11. اختيار الخوارزمية

تناسب DQN الأفعال المنفصلة، بينما تناسب DDPG وTD3 وSAC التحكم المستمر، ويمكن استخدام PPO في الحالتين.

12. Battery Scheduling as an MDP

State
SOC, load, price, PV, hour, forecasts.

Action
Charge or discharge power.

Transition
SOC dynamics and next-hour data.

Reward
Negative cost, degradation, and violations.

12. جدولة البطارية كـ MDP

تتكون الحالة من SOC والحمل والسعر والزمن، والفعل من قدرة البطارية، والمكافأة من الكلفة والتآكل والعقوبات.

13. Battery Dynamics

$$\mathrm{SOC}_{t+1} = \mathrm{SOC}_t + \frac{\eta_c P_{c,t}\,\Delta t}{E_{\max}} - \frac{P_{d,t}\,\Delta t}{\eta_d E_{\max}}$$

$$\mathrm{SOC}_{\min} \le \mathrm{SOC}_t \le \mathrm{SOC}_{\max}$$

13. ديناميكا البطارية

تزداد SOC عند الشحن وتنخفض عند التفريغ مع مراعاة الكفاءة والسعة وحدود التشغيل.

14. Reward Design

$$r_t = -c_t P_{\mathrm{grid},t}\Delta t - \lambda_{\mathrm{deg}}|P_{b,t}| - \lambda_{\mathrm{viol}}\,\mathrm{Violation}_t$$

Reward scaling strongly changes the learned behavior. Weak penalties may be ignored; excessive penalties can make the policy too conservative.

14. تصميم المكافأة

تجمع المكافأة كلفة الطاقة وتآكل البطارية وعقوبات تجاوز القيود، ويجب موازنة معاملات هذه الحدود بعناية.

15. RL versus Optimization

Aspect	Optimization	Reinforcement learning
Model	Usually explicit	Can learn from interaction
Online use	Solve repeatedly	Fast policy inference
Guarantees	Possible under assumptions	Usually limited
Best use	Well-modeled constrained problems	Complex sequential uncertainty

15. مقارنة RL والتحسين

يوفر التحسين ضمانات أقوى عند توفر نموذج واضح، بينما قد يوفر RL قرارًا سريعًا بعد التدريب. ولا ينبغي اعتباره بديلًا تلقائيًا للتحسين.

16. Safe Reinforcement Learning

- Action clipping
- Safety filters
- Constrained MDPs
- Lagrangian penalties
- MPC safety layers
- Offline validation

In a critical grid, learned actions should pass deterministic feasibility and protection checks.

16. التعلم المعزز الآمن

يمكن استخدام قص الأفعال ومرشحات الأمان والقيود وطبقات MPC، ويجب التحقق من القرار قبل تطبيقه.

17. Power-System Applications

Microgrids
Storage, generation, and grid exchange.

Demand response
Shift flexible loads.

Voltage control
Coordinate inverter reactive power.

Electric vehicles
Charging and vehicle-to-grid.

Restoration
Sequential switching after faults.

Market bidding
Storage participation under uncertainty.

17. التطبيقات

تشمل الشبكات المصغرة والاستجابة للطلب والتحكم بالجهد والمركبات واستعادة الشبكة والأسواق.

🔬 Research Corner

Emerging directions: offline RL, safe RL, multi-agent systems, model-based RL, distributional RL, and hybrid optimization-RL controllers.

🔬 زاوية البحث العلمي

من الاتجاهات الحديثة التعلم دون اتصال، التعلم الآمن، الأنظمة متعددة الوكلاء، التعلم القائم على نموذج، والدمج مع التحسين.

🧠 Review Questions

1. Define an MDP for battery scheduling.
2. Derive the Q-learning update.
3. Why does DQN need replay memory?
4. When is DQN unsuitable?
5. How does reward design affect degradation?
6. Why are safety filters needed?

🧠 أسئلة المراجعة

1. عرّف MDP لجدولة البطارية.
2. اشتق تحديث Q-learning.
3. لماذا تحتاج DQN إلى ذاكرة خبرة؟
4. متى لا تناسب DQN المسألة؟
5. كيف تؤثر المكافأة في التآكل؟
6. لماذا نحتاج مرشحات الأمان؟

References | المراجع

المراجع العامة المستخدمة في هذا المقرر

[1] A. Gabash, *Flexible Optimal Operations of Energy Supply Networks with Renewable Energy Generation and Battery Storage*. Saarbrücken, Germany: Südwestdeutscher Verlag für Hochschulschriften, 2014.

[2] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 4th Global ed. Harlow, U.K.: Pearson Education Limited, 2022.

[3] A. Gabash, “Energy Market Transition and Climate Change: A Review of TSOs-DSOs C+++ Framework from 1800 to Present,” *Energies*, vol. 16, no. 17, Art. no. 6139, 2023, doi: 10.3390/en16176139.

[4] A. Gabash, *AI Applications in Electrical Power Systems*, Version 1.0, 2026. [Online]. Available: https://aoussgabash.com

Recommended citation:

Gabash, A. (2026). Reinforcement Learning in Energy Management. AI Applications in Electrical Power Systems, Lecture 12, Version 1.0. https://aoussgabash.com

المرجع المقترح:

أوس غباش، 2026، محاضرة 12، تطبيقات الذكاء الاصطناعي في أنظمة الطاقة الكهربائية، الإصدار 1.0، https://aoussgabash.com

© 2026 Dr.-Ing. Aouss Gabash. Educational use with attribution to the author and official website.